Versatile prioritization of candidate disease genes or other molecules with NetworkPrioritizer

Tim Kacprowski[1,2], Nadezhda T. Doncheva[1], Mario Albrecht[1,2]
[1]*Max Planck Institute for Informatics, Saarbrücken and*
[2]*University Medicine Greifswald, Greifswald, Germany*

tim.kacprowski@uni-greifswald.de

Abstract: The prioritization of candidate disease genes or other molecules is often based on heterogeneous data. Nevertheless, most prioritization methods do not allow for a straight-forward integration of the user's own input data. Therefore, we developed NetworkPrioritizer, a Cytoscape plugin that enables the integrative network-based prioritization of bio-molecules. Our versatile software tool computes a number of important centrality measures to rank nodes based on their relevance for the connectivity in weighted and unweighted networks. As further novelty, it provides different methods to aggregate and compare rankings. NetworkPrioritizer and its documentation are freely available at http://www.networkprioritizer.de

1 Introduction

To elucidate the genetic foundations of human diseases, it is crucial to identify genes that might predispose to or cause specific diseases. Computational prioritization methods exploit the available biomedical knowledge to rank candidate genes according to their disease relevance. Many methods integrate multiple data sources, e.g., gene expression, protein interactions, and overlapping disease characteristics [DoKA12]. Integrated information of biological relationships and interactions is often represented as network. The connections between known disease genes and the remaining genes in a network are of particular interest as they can point to new disease genes.

The majority of prioritization methods are available only as web services [TCNM10]. Since the latter require the upload of the user's input data, they are not well suited to analyze confidential data. Furthermore, most web services rely on pre-defined background data and do not allow the user to include own data, to control the data integration, or to modify the aggregation of multiple rankings [DoKA12, TCNM10]. Existing Cytoscape plugins for

prioritization are also subject to a number of limitations. For instance, cytoHubba [LCWC08] and GPEC [LeKw12] rank network nodes focused on their close neighborhood or the steady-state probability of a random walk with restart, respectively. Neither supports further analysis or aggregation of rankings. The plugin NetworkAnalyzer [ARSL08] and the Java application CentiBiN [JuKS06] feature a large set of centrality measures, but cannot compute them for a user-defined set of seed nodes or for weighted networks.

Here, we present NetworkPrioritizer [KaDA13], a novel Cytoscape plugin for the integrative network-based prioritization of candidate genes or other molecules. It comprises two main functionalities. First, it facilitates the estimation of the relevance of network nodes, e.g., candidate genes, with regard to a set of seed nodes, e.g., known disease genes. Second, the plugin allows for the user-guided aggregation and comparison of multiple rankings.

2 Software Features

NetworkPrioritizer ranks nodes based on their relevance for the network connectivity, estimated by a number of centrality measures based on shortest paths and random walks (see web site). Closeness quantifies the distance of a node to the rest of the network. Betweenness measures the influence of a node on the paths connecting other nodes. These measures interpret all networks as undirected. NetworkPrioritizer can handle unweighted and weighted networks with user-adjustable effect of the edge weights on the computed centralities [OpAS10]. A particular feature of NetworkPrioritizer is the computation of the centrality measures with regard to a set of seed nodes.

In contrast to other prioritization tools, NetworkPrioritizer offers multiple methods to aggregate and compare multiple primary rankings. Weighted Borda Fuse (WBF) is a generalization of the popular Borda count [Saar99] and essentially ranks nodes according to their weighted mean rank in the primary rankings. Weighted AddScore Fuse (WASF) calculates the weighted sum of scores for each node in the primary rankings and awards a higher rank the larger this sum is. WBF and WASF can be used to identify candidate genes that attain high ranks in all primary rankings. If the primary rankings are based on scores on the same scale, WASF is more distinctive and thus more accurate than WBF. MaxRank Fuse (MRF) assigns each node the highest rank achieved in any primary ranking. MRF can quickly identify candidates with a high rank in at least one primary ranking. Furthermore, NetworkPrioritizer provides two widely used measures of ranking distance, the Spearman footrule and the Kendall tau [DKNS01]. The Spearman footrule is the sum of the rank differences of all nodes in two compared rankings. The Kendall tau is the number of nodes with different ranks.

3 Benchmark and Case Study

We evaluated the performance of NetworkPrioritizer in a comprehensive benchmark on data described in [ScLA10]. Briefly, this data contains artificial quantitative trait loci for 99 diseases extracted from OMIM [HSAB05] that have at least three known associated disease proteins. Disease-specific protein-protein interaction networks were extracted from BioMyn [RaLA12] and complemented by functional similarity links inferred from FunSimMat [ScLA10]. The receiver operating characteristics (ROCs) computed in a leave-one-out cross-validation are shown in Fig. 1. The shortest path-based centrality measures performed best, reaching areas under the ROC curve (AUCs) of up to 0.90. Notably, aggregating the primary rankings further improved the performance and resulted in AUCs from 0.87 to 0.92 (superior to the AUC of 0.85 achieved by MedSim on the same data).

As a case study, we applied NetworkPrioritizer to a Crohn's Disease (CD) related protein network. Protein-protein interactions and functional similarity links were compiled from BioMyn and FunSimMat, respectively, for proteins encoded by genes in CD-associated loci [FMBW10]. Proteins associated with inflammatory bowel disease (IBD), or CD as a subtype of IBD, were used as seed nodes (see web site). The 10 top-ranked proteins (HLA-B, SMAD3, CCL2, NOTCH1, STAT5A/B, HLA-A2/26/66, BECN1) function in the 'immune system process', 'response to stress', 'signal transduction', and 'homeostatic process' according to their Gene Ontology annotation. Since these processes are closely related to IBD [ZhLi12], the proteins are promising candidates for further experimental studies.

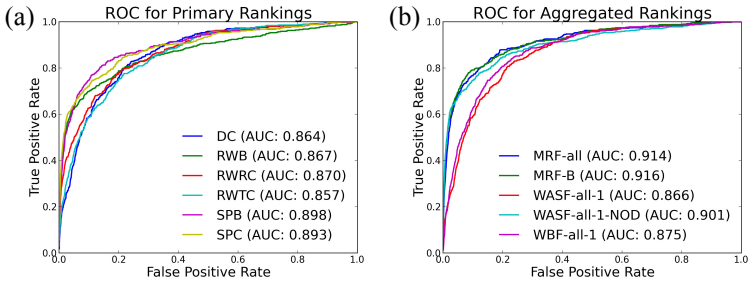

Fig. 1: (a) ROC for degree centrality (DC), random walk betweenness (RWB), random walk receiver closeness (RWRC), random walk transmitter closeness (RWRC), shortest path betweenness (SPB), and shortest path closeness (SPB). **(b)** ROC for the aggregation of all primary rankings using MRF (MRF-all), the aggregation of SPB and RWB (MRF-B), the aggregation of all rankings (equally weighted) using WASF (WASF-all-1), the aggregation of all rankings but DC (equally weighted) using WASF (WAS-all-1-NOD), and the aggregation of all rankings (equally weighted) using WBF (WBF-all-1).

4 Conclusions

NetworkPrioritizer enables the ranking of individual network nodes based on their relevance for connecting a set of seed nodes to the rest of the network. The Cytoscape plugin computes centrality measures for unweighted and weighted networks and, as a particular novelty, provides rank aggregation methods and ranking distance calculations. Its versatility makes NetworkPrioritizer a very useful tool for integrative network-based prioritization of candidate disease genes and proteins or other molecules.

References

[ARSL08] Yassen Assenov, Fidel Ramírez, Sven-Eric Schelhorn, et al. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–4, 2008

[DKNS01] Cynthia Dwork, Ravi Kumar, Moni Naor, et al. Rank aggregation methods for the web. *Proceedings of the 10th International Conference on World Wide Web - WWW*, 613–22, 2001

[DoKA12] Nadezhda T Doncheva, Tim Kacprowski and Mario Albrecht. Recent approaches to the prioritization of candidate disease genes. *Wiley Interdiscip Rev Syst Biol Med*, 4(5):429–42, 2012

[FMBW10] Andre Franke, Dermot PB McGovern, Jeffrey C Barrett, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*, 42(12):1118–25, 2010

[HSAB05] Ada Hamosh, Alan F Scott, Joanna S Amberger, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33:D514–7, 2005

[JuKS06] Björn H Junker, Dirk Koschützki and Falk Schreiber. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, 7:1–7, 2006

[KaDA13] Tim Kacprowski, Nadezhda T Doncheva and Mario Albrecht. NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, 29(11):1471–3, 2013

[LCWC08] Chung-Yen Lin, Chia-Hao Chin, Hsin-Hung Wu, et al. Hubba: hub objects analyzer – a framework of interactome hubs identification for network biology. *Nucleic Acids Res*, 36:W438–43, 2008

[LeKw12] Duc-Hau Le and Yung-keun Kwon. GPEC: a Cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection. *Comput Biol Chem*, 37:17–23, 2012

[OpAS10] Tore Opsahl, Filip Agneessens and John Skvoretz. Node centrality in weighted networks: generalizing degree and shortest paths. *Soc Networks*, 32(3):245–51, 2010

[RaLA12] Fidel Ramírez, Glenn Lawyer and Mario Albrecht. Novel search method for the discovery of functional relationships. *Bioinformatics*, 28(2):269–76, 2012

[Saar99] Donald G Saari. Explaining all three-alternative voting outcomes. *J Econ Theory*, 87(2):313–55, 1999

[ScLA10] Andreas Schlicker, Thomas Lengauer and Mario Albrecht. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*, 26(18):i561–7, 2010

[TCNM10] Léon-Charles Tranchevent, Francisco Bonachela Capdevila, Daniela Nitsch, et al. A guide to web tools to prioritize candidate genes. *Brief Bioinform*, 12(1):22–32, 2010

[ZhLi12] Hong Zhu and Y Robert Li. Oxidative stress and redox signaling mechanisms of inflammatory bowel disease: updated experimental and clinical evidence. *Exp Biol Med*, 237(5):474–80, 2012